Concept Drift Estimation with Graphical Models

Luigi Riso [1] *

Marco Guerzoni[2]

[1] Università Cattolica del Sacro Cuore
[2] DEMS, University of Milan-Bicocca, and BETA, Strasbourg University

September 21, 2022

Abstract

This paper deals with the issue of concept-drift in machine learning in the context of high dimensional problems. In contrast to previous concept drift detection methods, this application does not depend on the machine learning model in use for a specific target variable, but rather, it attempts to assess the concept drift as an independent characteristic of the evolution of a dataset. This major achievement enables data to be tested for the presence of drift, independently of the specific problem at hand. This is extremely useful when the same dataset is utilized for different classifications simultaneously, as it is often the case in a business environment. Moreover, unlike previous approaches, this method does not require the re-testing of each new model; a strategy which could prove expensive in computational terms. The fundamental intention of this work is to make use of graphical models to elicit the visible structure of data and represent it as a network. Specifically, we investigate how a graphical model evolves by looking at the creation of new links, and the disappearance of existing ones, in different time periods. We perform this task in four steps. We compute the adjacency matrix of a graph in each period, we apply a function that maps each possible state of the adjacency matrix over time into a transition matrix. We use the information in the transition matrix to produce a metric to estimate the presence of a drift in the data. Eventually, we evaluate this method with both three real-world datasets and a synthetic one.

Keywords: Drift Estimation, Graphical Models, Unsupervised Learning, Bayesian Logistic Regression

1 Introduction

Over the last several decades, the use of machine learning has become widespread across different industries due to both the increased availability of digitized information and improvements in algorithms. In particular, machine learning has become a standard tool for predicting key information in various organizational processes such as individual and corporate risk default, fraudulent claims, customer churn, machine failures [Nuccio and Guerzoni, 2019], and even COVID-19 variants specifications [Hussain et al., 2021]. The assessment of model uncertainty within a supervised machine learning exercise is based on testing its accuracy on a validation-set whose observations have not been used in the model training. This practice allows for flexibility in the choice of model and averts the risk of over-fitting. However, this process relies on the assumption that the data generating structure is common to both the test-set and future observations. While this assumption is rarely debatable in relation to physical processes, social processes are subject to change over time and so a model trained on past data might suffer a deterioration in its predictive power [Gama et al., 2014]. This phenomenon, which is known as concept or model drift, describes a situation in which there exists a hidden context of the data-generative structure, i.e any effect of the outcome variable which is not captured by the model features and which changes abruptly, incrementally, or periodically, over time [Widmer and Kubat, 1996, Webb et al., 2016]. Scholars have addressed this issue and developed a battery of techniques for concept drift detection. As reviewed in Klinkenberg and Joachims [2000] and Elwell and Polikar [2011], traditional techniques typically rely on

*Corresponding author: Luigi Riso, Università Cattolica del Sacro Cuore, luigi.riso@unicatt.it, Largo Agostino Gemelli, 1, 20123, Milano, MI.

adopting different time windows or size of the training data [Klinkenberg and Renz, 1998, Gama et al., 2013] or on explaining how the weights of different features of a predicted outcome change over time [Klinkenberg and Renz, 1998, Taylor et al., 1997, Klinkenberg, 2004, Zhao et al., 2020]. A recent review [Althabiti and Abdullah, 2020] surveys methods that can also deal with model update with stream data [Bose et al., 2011]. However, most of these techniques make use of a statistical comparison of the changes in classification error and, from this evidence, they deduce the presence of concept drift [Widmer and Kubat, 1996]. In this paper, we approach this problem from a different angle. We apply graphical models[Lauritzen, 1996] to elicit the visible structure of the data and we estimate its changes over time. Thus, as distinct from previous concept drift detection methods, this application does not depend on the machine learning model in use, but rather, it assesses concept drift as an independent characteristic of the data. For instance, Barros and Santos [2018]'s test of 14 different detector methods had to be performed using just two specific classifiers and nothing could be inferred from either one. Thus, the method presented in this paper releases drift detection from specific classifiers. This approach belongs to the body of work that uses graphical models to represent concept drift [Borchani et al., 2015, Cabañas et al., 2018], although it departs from this in that we attempt to model the entire joint distribution of variables instead of focusing solely on the relation with a specific label variable to be predicted. Eventually, we propose an algorithm capable of computing the drift and apply it to real-world datasets.

2 Problem statement and related literature

2.1 Problem statement

In this paper, we aim to estimate the concept drift of a dataset. If we take a dataset $\mathbf{X}_t = (X_{1,t}, \ldots, X_{p,t})$ observed in different points in time, $t = 1, 2, \ldots, T$ and define its concept at time t as the joint distribution $P(\mathbf{X}_t)$. A concept drift exists if $P(\boldsymbol{X}_t) \neq P(\mathbf{X}_{k \neq t})$. The aim of the paper is to provide an estimation β_t of the magnitude of the concept drift from $P(\mathbf{X}_{t=1})$ to $P(\mathbf{X}_{k \neq t})$ with $t = 1, 2, \ldots, T$.

2.2 Related literature

The definition of "concept" adopted in this paper is one of the two provided in the related literature. According to Webb et al. [2016], a narrow definition of concept implies a specific, supervised, exercise in which one variable is not a feature but a label to be classified. Thus, for the specific classification problem of the variable C, the concept is defined as the joint distribution $P(X_{1,t}, \ldots, X_{p-1,t}|C_t)$, where the variable p_{th} is considered a label to be classified, C. Alternatively, for cases beyond classification, as for instance, in unsupervised learning, the broad definition of concept is the joint distribution $P(X_{1,t}, t, \ldots, X_{p-1,t}, C_t)$. Borchani et al. [2015] define the concept drift as when there is a change in the joint distribution over time $P(\cdot)_t \neq P(\cdot)_{k \neq t}$. This general definition applies to both cases of concept drift introduced by Webb et al. [2016]. However, when Borchani et al. [2015] discuss available tools in concept drift detection, and distinguish between real and virtual concept drift, they clearly refer to the supervised case only. To be exact, the real concept drift is when the joint distribution $P(X_{1,t}, \ldots, X_{p-1,t}|C_t)$ changes over time and the virtual concept drift i.e. $P(X_{1,t}, \ldots, X_{p-1,t})$ changes, but $P(X_{1,t}, \ldots, X_{p-1,t}|C_t)$ remains constant. The former case describes the situation where the drift changes the relationship between variables and the label to be predicted, while in the latter situation, the drift changes the relationships among variables, but that between variables and the label remains unaltered. In both cases, the elicitation of a drift always refers to the concept drift for a specific classification problem of the label C.

The idea behind this characterization is that the focus is on detecting the change, which eventually impacts upon the accuracy of the specific classification model at hand. For this reason, all of the recent contributions reviewed in Klinkenberg and Joachims [2000], Elwell and Polikar [2011], Althabiti and Abdullah [2020], rely on this definition of concept drift. For its estimation, they measure, over time, the increase in the classification error for the target variable C.

However, this useful simplification does not come without a cost, since the measure of the drift is wholly specific to both the prediction of the label variable and the supervised model adopted: when the same dataset is used for different classification problems, the researcher needs to evaluate different concept drifts for different problems.

For this reason, while other methods are specific to a certain problem, in this paper, we approach a situation rarely discussed in related literature but encompassed in Webb et al. [2016]'s definition of the most general category of concept drift. Specifically, any change in the joint distribution of the random variables in a dataset $P(X_1, t, \ldots, X_{p-1,t}, C_t)$ regardless of the specific problem under consideration

and a related target variable. In other words, we consider variable C_t as p_{th} variable in the dataset $\mathbf{X}_t = (X_{1,t}, \ldots, X_{p-1,t}, X_{p,t})$ and we are interesting in analyzing whether the joint distribution changes over time $P(\mathbf{X}_t) \neq P(\mathbf{X}_{k \neq t})$.

Therefore, both in the approach and in the notation we follow Borchani et al. [2015] by describing the problem of drift detection as a comparison of different distributions overtime, but we radically depart from them in that we consider $P(\mathbf{X}_t)$ without making a distinction between variables and labels. Our approach, which does not delimit the computation of drift solely for a specific classification model, makes it possible to measure how the joint probability $P(\mathbf{X}_t)$ drifts over time and has the following advantages. First, we address the drift in the dataset without imposing any a-priori classification problem. This is useful in many contexts, such as in a business environment in which the same data are employed for different exercises, both supervised and unsupervised. Second, and pragmatically, the theoretical case of virtual drift is very unlikely to occur in datasets representing highly complex interactions. Third, we do not make the simplistic assumption, as in [Borchani et al., 2015, Cabañas et al., 2018], that variables in X are mutually independent[1].

3 A measure of dynamic stability as proxy for the concept drift

The task of estimating a measurement of the absence of a drift from time t to T, β_t, for the dataset $\mathbf{X}_t$ consisting of p variables observed at different points in $t = 1, 2, \ldots, T$ (hours, days, weeks, months, years, ...) require the following steps: first, we compute an approximation of the joint probability $P(\mathbf{X}_t)$ over T periods whit the support of the graphical models, second, we present a method by which to compare $P(\mathbf{X}_t)$ over time t, with the use of a transition matrix and, finally, we provide an estimator β_t which can be considered as an empirical proxy for concept drift.

3.1 Graphical models and the estimation of the joint distribution

The first step in our approach requires the encoding of the joint probability $P(\mathbf{X_t})$ into a graphical model, which maps the conditional dependence relationships in a undirected graph $\mathbf{G}$. $\mathbf{G} = (V, E)$ is a mathematical object where V is a finite set of nodes $\{V_1, \ldots, V_p\}$, with a one-to-one correspondence with the p random variables $\{X_1, \ldots, X_p\}$, $E \subset V \times V$ is a subset of ordered couples of V representing the conditional dependence between any two nodes V_i and V_j, mapping the p variables [Jordan et al., 2004]. We encode, on the graph, a very simple approximation of the joint distribution: if a link between two nodes is absent, the two variables represented by the nodes are independent conditional upon the dependence of the remaining variables, and are not independent otherwise [Lauritzen, 1996]. Therefore, we do not take the magnitude of these links into account. Moreover, to reduce the space taken up by possible graphs, which in the case of p variables can add up to $2^{p(p-1)/2}$, we make the further assumption that the approximation of the joint distribution is encoded in a "tree" or a "forest". A tree is 'a decomposable graph with cliques of size two (or less) and such that any two non-adjacent nodes are separated by a set of (at most) size one' [Carota et al., 2014]. In order words, the graphs does not display any triangulation as in Figure 1. A forest is a set of disconnected trees.

In this paper, we derive the graph with the approach adopted by Edwards et al. [2010], an extension of the Chow and Liu [1968] algorithm which can be applied with each type of dataset variable composition (discrete, continuous or mixed). Furthermore, Edwards et al. [2010] suggest penalizing the tree/forest maximum likelihood with the AIC or BIC criterion[2], in order to avoid the inclusion of links not supported by data[3].

[1] Actually [Cabañas et al., 2018] suggest that this assumption can be easily removed. However, the computation can become really cumbersome, while, as explained below, our approach enables the algorithm to deal with very large datasets

[2] $I^{AIC} = I(x_i, x_j) - 2k_{x_i,x_j}$ or $I^{BIC} = I(x_i, x_j) - \log(n)k_{x_i,x_j}$ where k_{x_i,x_j} are the degrees of freedom associated with the pair of variables, that are defined according to the nature of the variables involved [Akaike, 1974, Schwarz, 1978]

[3] Details about graphical model can be found in [Lauritzen, 1996]'s textbook and they are also discussed elsewhere by the authors [Edwards et al., 2010]

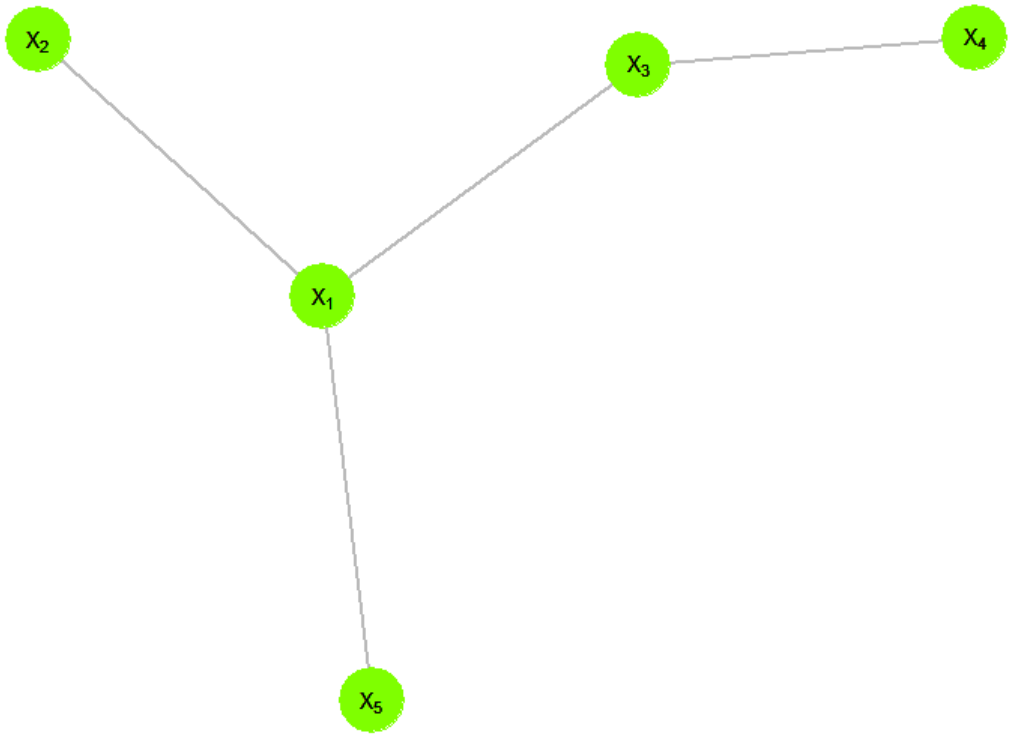

Figure 1: Example min BIC tree, from a dataset $\mathbf{X}$ with $p = 5$ variables and N observation at time t

Graph $\mathbf{G}$ approximates the joint probability of the entire dataset $\mathbf{X}$, observed at a generic time t (as in Figure1) and, as with graph, can be always represented by its adjacency matrix AM_t, which is a symmetric matrix, with dimension $(V \times V)$, in which each element takes value of 1 if an edge exists between two of the p variables, and zero otherwise. Elements in the main diagonal are zeros, since self-loops are not allowed [Edwards et al., 2010]. For example the, the Adjacency Matrix of the graph in Figure 1 is equal to:

$$AM_t = \begin{array}{c} \begin{array}{ccccc} X_1 & X_2 & X_3 & X_4 & X_5 \end{array} \\ \left(\begin{array}{ccccc} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{array} \right) \end{array} \begin{array}{c} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{array}$$

In the approach proposed here, the detection of changes over time in AM_t, is the cornerstone of the strategy to measure the concept drift.

3.2 Transition Matrix Processes

A second step in the computation of the drift is the detection of changes in the adjacency matrices AM_t with $t = 1 \dots T$. We introduce a function that maps all the observed changes in $AM_{t\in(1,T)}$, over time, into a transition matrix $TM_T = f(AM_{t\in(1,T)})$, denoted TM_T, of dimension $(V \times V)$. Its generic element $w_{i,j}$ registers, for each couple of node/variable i and j, the possible sequence of 0 and 1 occurred in T periods.

Definition 3.1 (Transition Matrix Process)**.** Given a temporal period of observation $t = 1, 2, \dots, T$, the transition matrix process TM_T:

$$TM_T = \sum_{t=1}^{T} 2^{(T-t)} AM_t, \tag{1}$$

records any possible states of the AM_1,AM_2,...,AM_T.

In order to illustrate definition 3.1, the following paragraph depicts the transition matrix process up to $T = 3$ and, thereafter, generalizes for T periods. As a starting point, in $t = 1$ the transition matrix TM_1 is equal to the adjacency matrix $AM_{t=1}$, where $w_{i,j;1} = 0$ means that the i-node and j-node are not connected, while when $w_{i,j;1} = 1$ means that the V_i node and V_j node are connected. At $t = 2$ existing links may or may not persist, while non-existing links may or may not appear. From Equation 1,

$$TM_2 = 2 \times AM_1 + AM_2 \tag{2}$$

Thus, TM_2 maps any possible evolution of the connections between the V_i and V_j nodes, with $w_{i,j;2}$ able to take values $\{0, 1, 2, 3\}$. When V_i and V_j are never connected, that is $AM_{i,j;t=1} = AM_{i,j;t=2} = 0$, then $w_{i,j;2} = 0$. If V_i and V_j stay connected, that is $AM_{i,j;t=1} = AM_{i,j;t=2} = 1$, then $w_{i,j;2} = 3$. For $AM_{i,j}$ changing from 0 in $t = 1$ to 1 in $t = 2$ we have $w_{i,j;2} = 1$, while $w_{i,j;2} = 2$ for $AM_{i,j}$ changing from 1 in $t = 1$ to 0 in $t = 2$. Table 1 shows the possible scenarios at $T = 2$.

Table 1: All possible AM_t values for two nodes V_i and V_j and the resulting $w_{i,j}$ in TM_T function for $T = 2$

$AM_{i,j;1}$	$AM_{i,j;2}$	$TM_{i,j;2} = w_{i,j;2}$
0	0	**0**
0	1	**1**
1	0	**2**
1	1	**3**

At time $T = 3$ the possible evolution of AM_t can be described by 8 values:

$$TM_3 = 2^2 \times AM_1 + 2^1 \times AM_2 + 2^0 \times AM_3 \tag{3}$$

Table 2 summarizes all possible combinations between two generic nodes (V_i, V_j) of binary values of the $AM_{i,j;1}, AM_{i,j;2}, AM_{i,j;3}$, mapped into $TM_{i,j;3}$. Generally, for time T we can derive Eq. 1:

$$\begin{aligned}
TM_2 &= 2 \times AM_1 + AM_2 \\
TM_3 &= 2 \times TM_2 + AM_3 \\
TM_3 &= 2 \times (2 \times AM_1 + AM_2) + AM_3 \\
TM_3 &= 2^2 \times AM_1 + 2^1 \times AM_2 + 2^0 \times AM_3 \\
TM_3 &= \sum_{t=1}^{3} 2^{(3-t)} AM_t \\
&\dots \\
TM_T &= \sum_{t=1}^{T} 2^{(T-t)} AM_t
\end{aligned} \tag{4}$$

Table 2: All possible AM_t values for two nodes V_i and V_j and the resulting $w_{i,j}$ in TM_T function for $T = 3$

$AM_{i,j;1}$	$AM_{i,j;2}$	$AM_{i,j;3}$	$TM_{i,j;3} = w_{i,j;3}$
0	0	0	**0**
0	0	1	**1**
0	1	0	**2**
0	1	1	**3**
1	0	0	**4**
1	0	1	**5**
1	1	0	**6**
1	1	1	**7**

In general, the value of the generic element $w_{i,j;t} \in \mathcal{W}_t \subset \mathbb{N}_0$ of TM_t, can be considered a realization of a discrete random variable $\mathcal{W}_t$ with density $f(w_{i,j;t})$:

$$f(w_{i,j;t}) = P(\mathcal{W}_{i,j;t} = w_{i,j;t}), \quad t = 2, \ldots, T \tag{5}$$

Thus, $TM_{i,j;T}$ represents the evolution of the connection between the i-node and the j-node at time T, for each node V. The numerosity of the set $\mathcal{W}_T = \{0, 1, 2.., 2^T - 1\}$ is 2^T.

3.3 From the transition matrix to stability

The main idea of the paper is to measure the appearance or disappearance of connections between nodes as a proxy for model drift. For this reason, we are especially interested in two specific levels which describe the absence of change. That describing the state of the word in which a connection between two nodes never exist, i.e. $AM_{i,j;t} = 0 \ \forall \ t$ and that which describes a stable connection over time, i.e. $AM_{i,j;t} = 1 \ \forall \ t$. In the case where $T = 3$, the two cases map into $w_{i,j;3} = 0$ and $w_{i,j;3} = 7$, as showed in Table 2. In general for a generic t, we have a stability of connections when connections are always absent, with $w_{i,j;t} = 0$, or always existing, with $w_{i,j;t} = 2^t - 1$. Thus, for each value of t the transition matrix introduces a partition within the set of all possible $\mathcal{V}$ connections:

$$\mathcal{V} = \frac{V(V-1)}{2}$$

between the V nodes in the undirected, graph as shown in Figure 2.

Indeed, each transition at time t generates a subsequent partition, $\mathcal{V}$, dividing the possible connections in two groups: stable ones, with $w_{i,j;t} = 0$ or $w_{i,j;t} = 2^t - 1$ and the others. This process, encoded in the transition matrix can be seen as a particular case of a tail-free processes such as proposed by Jara and Hanson [2011]. Consider a sequence $\mathcal{T}_1 = \{\mathcal{V}\}$, $\mathcal{T}_2 = \{A_0, A_1\}$, $\mathcal{T}_3 = \{A_{00}, A_{01}, A_1\}$, and so on, of measurable partitions of the $\mathcal{V}$ elements, obtained by slitting every set in the preceding partition into two new sets for the node on the left and maintain the same node for the others.

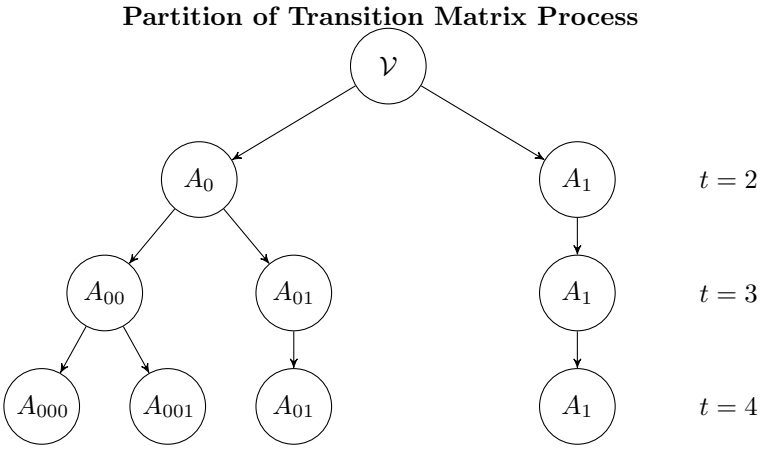

Figure 2: Representation of Transition Matrix process with Tail-free processes

At each time t, we can partition the connections between stable and unstable ones. Figure 2 shows a tree diagram representing the distribution of mass over time $\mathcal{V} = A_0 \cup A_1 = (A_{00} \cup A_{01}) \cup A_1$ of the elements of the transition matrix at each time. At time $t = 2$, A_0 contains elements $w_{i,j;2} = \{0, 3\}$, indicating stable connections, while A_1 includes the remaining ones. At the subsequent period, A_0 is partitioned between A_{00} and A_{01}. The former includes stable connections, $w_{i,j;3} = \{0, 7\}$, while the latter ones, $w_{i,j;3} = \{1, 6\}$, prove unstable at time $t = 3$, like those already present in A_1. This partition process, registered in the transition matrix TM_t, can generate a simplified version of transition matrix Q_t with values :

$$Q_t = \begin{cases} q_{i,j;t} = 1 & \text{if } w_{i,j;t} = 0 \vee w_{i,j;t} = 2^t - 1 \\ q_{i,j;t} = 0 & \text{otherwise} \end{cases}, \quad t = 2, \ldots, T \tag{6}$$

Thus, Q_t indicates, for each pair of variables, whether the status of their dependence is consistent, over time, with $q_{i,j;t} = 1$ or inconsistent with $q_{i,j;t} = 0$. We define Y_t as the half-vectorization of Q_t without the main diagonal elements, $Y_t = vech(Q_t) \backslash diag(Q_t)$ with length $\mathcal{V}$. As TM_1 tallies with AM_1, the partition starts from $t = 2$, which means that $T - 1$ transition matrices turn out to be associated to

T adjacency matrices, as explained in Figure 3:

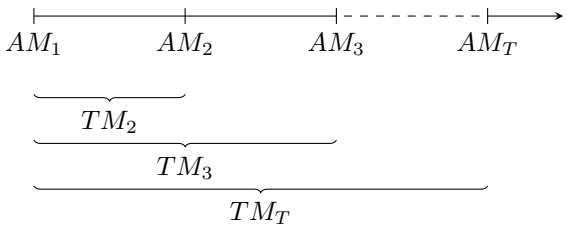

Figure 3: From the adjacency matrices to the transition matrices

It is worth noting that the structure of the transition matrix process turns out to be dependent on the min BIC spanning tree/forest at time $t = 1$. For each period $t = 2, \ldots, T$, the normalized partition of $\mathcal{V}$ is given by:

$$\lambda_t = \frac{\sum_{i \in t} Y_i}{\mathcal{V}}, \quad t = 2, \ldots, T \tag{7}$$

Therefore, we pool together by rows binding the $T - 1$ transition periods and define *Stability*, the resulting variable Y with length $\mathcal{V} \times (T - 1)$:

$$Q_t = \begin{bmatrix} q_{1,1;t} & q_{1,2;t} & \cdots & q_{1,V;t} \\ q_{2,1;t} & q_{2,2;t} & \cdots & q_{1,V;t} \\ \vdots & \vdots & \ddots & \vdots \\ q_{V,1;t} & q_{V,2;t} & \cdots & q_{V,V;t} \end{bmatrix} \xrightarrow[without\ the\ main\ diagonal]{Half-vectorization} Y_t = \begin{bmatrix} q_{1,2;t} \\ q_{1,3;t} \\ \vdots \\ q_{V-1,V;t} \end{bmatrix} \xrightarrow[different\ t]{Pooling} Y = \begin{bmatrix} Y_2 \\ Y_3 \\ \vdots \\ Y_T \end{bmatrix}$$

We also take the half-vectorization of the values TM_t, $t = 2, \ldots, T$, without the elements of the main diagonal: $W_t = vech(TM_t) \backslash diag(TM_t)$, a variable with length $\mathcal{V}$. Similarly we pool together the $T - 1$ W_t and obtain the variable W with length $\mathcal{V} \times (T - 1)$:

$$TM_t = \begin{bmatrix} w_{1,1;t} & w_{1,2;t} & \cdots & w_{1,V;t} \\ w_{2,1;t} & w_{2,2;t} & \cdots & w_{1,V;t} \\ \vdots & \vdots & \ddots & \vdots \\ w_{V,1;t} & w_{V,2;t} & \cdots & w_{V,V;t} \end{bmatrix} \xrightarrow[without\ the\ main\ diagonal]{Half-vectorization} W_t = \begin{bmatrix} w_{1,2;t} \\ w_{1,3;t} \\ \vdots \\ w_{V-1,V;t} \end{bmatrix} \xrightarrow[different\ t]{Pooling} W = \begin{bmatrix} W_2 \\ W_3 \\ \vdots \\ W_T \end{bmatrix}$$

Stability is the cornerstone of our strategy and in the next paragraph we explain how we can use it together with W to estimate an empirical measure of model drift.

3.4 Drift Estimation

In this section, we explain how can we use the *Stability* Y as a latent variable to estimate both the presence and the magnitude of the drift.
Consider the following variable with same length n, where $n = \mathcal{V} \times (T - 1)$:

- Y, as defined above
- W, as defined above
- T the corresponding time for each Y.

Then, we build a dataset with these variables and designate it $\mathbf{D}$. Note that by definition, the observations of $\mathbf{D}$ is exchangeable since we have built $\mathbf{D}$ respecting the temporal period of the adjacent matrices, thus:

$$P(\mathbf{D}_1, \ldots, \mathbf{D}_n) = P(\mathbf{D}_{\sigma(1)}, \ldots, \mathbf{D}_{\sigma(n)})$$

for all $n \geq 1$ and all permutations σ of $(1, \ldots, n)$. In other words, the order of appearance of the observations does not matter in terms of their joint distribution. In order to exploit this property, we

implement a Bayesian Regression Model using the dataset **D** [Albert and Hu, 2019]. One advantage of a Bayes perspective is the opportunity to consider the context of the analysis with the support of prior distribution. This approach functions to discern the situations in which the data generating structure is common to the test-set and future observations as a physical process, with respect to a social process that can change very quickly and abruptly over time [Box, 1980]. Furthermore, as Gelman et al. [2008] suggest, non-identifiability is a common problem in classical logistic regression. In addition to the problem of collinearity, typical in linear regression, discrete-data regression can also become unstable due to separation, which arises when a linear combination of the predictors can perfectly predict the outcome [Albert and Anderson, 1984, Lesaffre and Albert, 1989]. Separation is surprisingly common in applied logistic regression especially with binary predictors [Zorn, 2005]. In this context, Bayesian inference is a valid alternative approach to obtaining stable coefficients [Gelman et al., 2008].

Let θ_i be the probability of a realization of $Y_i = 1$ with odds $\frac{\theta_i}{1-\theta_i}$. Thus the dichotomous variable Y can be described by a Bernoulli distribution defined as follows:

$$Y_i|\theta_i \overset{ind}{\sim} Bern(\theta_i), \quad i = 1, .., n$$

Now, consider a logistic regression model[4], in which the logit of the probability θ_i, or the log of its odd, is a linear function of some predictors $\mathbf{x}_i$:

$$Logistic(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \sum_{j=1}^{2^T} \beta_j \mathbf{x}_{j,i} \tag{8}$$

For the case under exam, the j predictors are 2^T, with T, time of the of Y and the different levels of the variable W. Since W has 2^T levels, we regress $2^T - 1$ dummy variable and keep $W = \{0\}$ as the reference category:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i} \tag{9}$$

By construction, the intercept of this model β_0 can be interpreted as the baseline risk for *Stability*. A high β_0 suggests that the underlying graphical model does not change much over time. β_t captures the effect of the drift over time. It can be shown that *Stability* is slightly decreasing over time and, thus β_t defines the speed of convergence towards the absence of *Stability*, or alternatively the presence of the drift. Finally, since the variable Y takes value 1 for $W = \{0, 2^{t-1}\}$, the coefficient β_{2^T-1}, i.e. the coefficient for $W = \{2^{T-1}\}$ with reference $W = \{0\}$ captures component of *Stability* which originates in the persistence of existing connections, rather than in the continuing absence of connections.

The computation is straightforward: by rearranging the logistic regression in Equation 9, it is possible to express it as a nonlinear equation for the probability of success θ_i :

$$\begin{aligned}
\log\left(\frac{\theta_i}{1-\theta_i}\right) &= \beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i} \\
\frac{\theta_i}{1-\theta_i} &= \exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\} \\
\theta_i &= \frac{\exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\}}{1 + \exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\}}
\end{aligned} \tag{10}$$

From the Equation 10, we can define the likelihood of the sequence of Y_i over data set of n subjects:

$$\begin{aligned}
p(Y_i|\beta_0, \beta_t, \ldots, \beta_{2^T-1}) = \prod_{i=1}^{n} \Bigg[&\left(\frac{\exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\}}{1 + \exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\}}\right)^{y_i} \\
&\left(1 - \frac{\exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\}}{1 + \exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\}}\right)^{(1-y_i)} \Bigg]
\end{aligned} \tag{11}$$

[4]The logistic regression seems the most natural way to describe this phenomenon. However, according to the type of expected drift, we could employ other functions, without loss of generalization.

The set of unknown parameters consists of $\beta_0, \beta_t, \ldots, \beta_{2^T-1}$. Generally, any prior distribution can be used, depending on the available prior information [Carlin and Louis, 2008]. As it is usually implemented (ibid.), if something is known about the likely values of the unknown parameters, this information can allow for the use of informative prior distribution. Alternatively, the use of non-informative prior is common when little is known about the coefficient values or where the goal is to exploit data-driven inference [Kass and Wasserman, 1996, Box, 1980]. Numerous proposals and divergent opinions exist as to the choice of prior for logistic regression. For example, Hanson et al. [2014] provide a simple Gaussian g-prior for logistic regression coefficients according to variants of the g-prior proposed by Rathbun and Fei [2006], Marin et al. [2007] and Bové and Held [2011], while Gelman et al. [2008] suggest standardizing non-binary covariates and then placing independent Cauchy priors on regression coefficients based on how covariates could reasonably affect the odds of the response. However, their insightful approach does not take into account correlations between the predictor variables [Hanson et al., 2014]. The selection of the prior is a key point of the Bayesian framework. Indeed the choice of an improper prior can lead to an improper posterior. According to Syversveen [1998],a way of overcoming problems with improper prior is to use proper approximations to improper priors. Examples are normal distributions with large variance or a uniform distribution on a compact set. The fundamental disadvantage of using the uniform distribution as our non-informative prior, is that uniform distribution is not invariant under reparametrization [Syversveen, 1998]. For these reasons, in this case, we suggest the use of the normal distribution, which is the most common distribution for establishing priors for logistic regression parameters [Cramer, 2002, Genkin et al., 2007]:

$$\begin{aligned} \beta_0 &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ \beta_t &\sim \mathcal{N}(\mu_t, \sigma_t^2) \\ \beta_j &\sim \mathcal{N}(\mu_j, \sigma_j^2), \quad j = 1, \ldots, 2^{T-1} \end{aligned} \tag{12}$$

The most common choice for μ is zero with σ^2 large enough to be considered non-informative in the range from $\sigma = 10$ to $\sigma = 100$ [Albert and Hu, 2019]. However, if we have sufficiently deep and detailed knowledge of the problem we are studying, the choice of informative priors will not change the innovative process proposed in this paper. In order to make this paper easier to read, we synthesize the variance vector of the priors with $\boldsymbol{\sigma}^2$, in this way we have $\boldsymbol{\sigma}^2 = (\sigma_0^2, \sigma_t^2, \ldots, \sigma_{2^T-1}^2)$, while with the vector $\boldsymbol{\mu}$ we define the mean of the priors $\boldsymbol{\mu} = (\mu_0, \mu_t, \ldots, \mu_{2^T-1})$, finally whit $\boldsymbol{\beta}$ we define the coefficient vector of the logistic regression $\boldsymbol{\beta} = (\beta_0, \beta_t, \ldots, \beta_{2^T-1})$. According to these specifications, the following hierarchical model emerges:

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} Bern(\theta) \\ \theta_i|\boldsymbol{\beta} &\sim Logistic(\boldsymbol{\beta}) \\ \boldsymbol{\beta} &\sim \mathcal{N}_{2^T+1}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)) \end{aligned}$$

[5]

The posterior distribution of $\boldsymbol{\beta}$ is extrapolated by combining likelihood Equation 11, with the priors in Equation 12 [O'brien and Dunson, 2004]:

$$\begin{aligned} p(Y_i|\beta_0, \beta_t, \ldots, \beta_{2^T-1}) = \prod_{i=1}^{n} \Bigg[& \left(\frac{\exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\}}{1 + \exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\}} \right)^{y_i} \\ & \left(1 - \frac{\exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\}}{1 + \exp\left\{\beta_0 + \beta_t \times t_i + \sum_{j=1}^{2^T-1} \beta_j w_{j,i}\right\}} \right)^{(1-y_i)} \Bigg] \times \\ & \times \prod_{j=0}^{2^T-1} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2} \left(\frac{\beta_j - \mu_j}{\sigma_j} \right)^2 \right\} \times \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left\{ -\frac{1}{2} \left(\frac{\beta_t - \mu_t}{\sigma_t} \right)^2 \right\} \end{aligned} \tag{13}$$

However, the regression parameters β_j, with $j = 1, \ldots, 2^{T-1}$, are of limited interest. In fact, our interest is captured by the parameter β_t which defines the evolution of the *Stability* over the time and

[5] which is the multivariate normal distribution $2^T + 1$-dimensional.

thus an estimation of the drift. Furthermore, the Bayesian approach gives us the opportunity to compute the prediction of the *Stability* over a specific time t [Gelman et al., 1995]. If $\widetilde{y_i}$ represents the number of similarity connection between n nodes at time t, then one would be interested in the posterior predictive distribution of the fraction $\widetilde{y_i}/n$ One represents this predictive density of $\widetilde{y_i}$ as:

$$f(\widetilde{Y_i} = \widetilde{y_i}|y) = \int p(\boldsymbol{\beta}|Y, \boldsymbol{\mu}, \boldsymbol{\sigma}) p(\widetilde{y_i}, \boldsymbol{\phi}) d\boldsymbol{\phi} \tag{14}$$

where $p(\boldsymbol{\beta}|Y, \boldsymbol{\mu}, \boldsymbol{\sigma})$ is the posterior density and $p(\widetilde{y_i}, \beta)$ is the Binomial sampling density of $\widetilde{y_i}$ conditional of regression vector $\boldsymbol{\phi} = (\beta_0, \beta_t)$ [Genkin et al., 2007, Albert and Hu, 2019]. Figure 4 represents the Bayesian graphical model of the drift estimation, which synthesizes the entire process from the adjacent matrices to the estimation of the Bayesian logistic parameters.

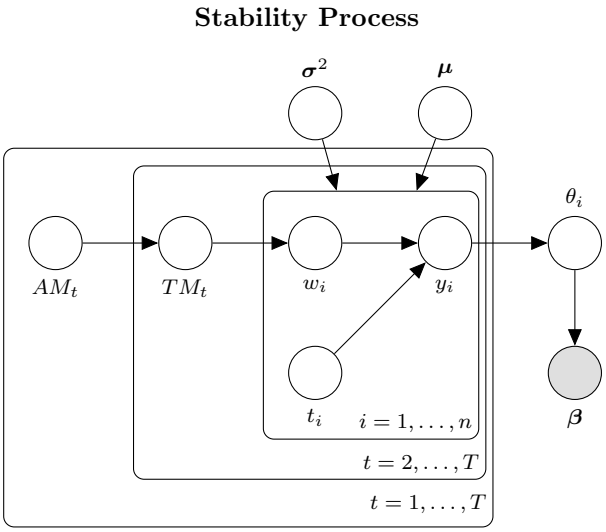

Figure 4: Bayesian Graphical Model of the *Stability*

4 Empirical experiments

As a testbed for this theoretical approach, we will apply the *Stability* index by using four different examples, three of which belong to real-world datasets, while the final example is a simulated dataset. For the first three, we know when the drift occurs from previous literature, while for the third we know by definition. The first two datasets, *ELEC2* and *Airlines*, are commonly used in various papers that evaluate the presence of drift [Baena-Garcıa et al., 2006, Webb et al., 2018], while the third dataset *Ozone Level* [Zhang and Fan, 2008] shows how the index *Stability* can be applied to high dimensional fields. The fourth dataset, *Harvard mixed gradual drift dataset*, belongs to a collection of datasets proposed by López Lobo [2020], in order to know beforehand where a drift occurs. A direct comparison with others' methods, applied in the first two datasets, is not straightforward since the present method evaluates the drift of the overall joint-probability structure and does not concern one label variable only. However, this method appears to be remarkably efficient at identifying when the drift occurs. The third dataset has not previously been employed for drift detection. However, Zhang and Fan [2008] suggest the presence of an incremental drift and the year on year prediction they need to calibrate the model in each period. The fourth dataset is a synthetic one and we precisely know when the drift occurs and we can evaluate that the *Stability* index spot the right moment. All in all, the choice of these datasets covers both abrupt and incremental changes as well as domains with few, or numerous, variables. For all examples, not having detailed knowledge about these datasets, we used non-informative priors for the logistic regression [Albert and Hu, 2019], that is from a normal distribution with with $\boldsymbol{\mu} = (0, 0, \ldots, 0)$ and $\boldsymbol{\sigma}^2 = (100, 100, \ldots, 100)$.

4.1 *ELEC2* dataset

The first example is *ELEC2* dataset [Harries, 1999], which is a benchmark for drift evaluation [Baena-Garcıa et al., 2006, Kuncheva and Plumpton, 2008, among many others]. This dataset is available on R and contains information about the Australian New South Wales (NSW) Electricity Market, consisting of

27552 records, dated from May 1996 to December 1998, each referring to a period of 30 minutes. These records have 5 fields: a binary class label and four covariates capturing different aspects of electricity supply and demand. In order to compute the empirical evolution of the drift over time, we group observations into one week periods. Thus, for each week we have a panel dataset of 5 variables with 336 observations for $T = 82$ periods. First, we realize a graphical model for each period t. Table 7 shows the label of the node and its corresponding variable. Figure 5 portrays the graphs for selected periods and shows that the structure of the graph changes over time. We thus expect the presence of a drift. Figure 6 depicts the evaluation of the *Stability* overtime. We can observe 5 moments of non-stationarity which correspond to the presence of the drift ($t = 8$, $t = 12$, $t = 14$, $t = 19$, and $t = 41$). The red dots are the ratio of stable relations among variables to all possible connections for a given period λ_t (Equation 7). The blue line is the estimation of Equation 13 with its related credible interval represented by the gray contour. In Table 3 are reported the posterior summaries for the regression parameters. The presence of the drift over time is given by the mean of $\beta_t = -0.3$, while the mean of $\beta_0 = 7.7$ identifies the stability within the dataset. Figure 6 shows also that in the last period of observation *Stability* is close to 0 and the credibility interval includes negative values, since the initial concept drifted a lot. However, this deterioration is slow and punctuated by precise moments in which the drift occurs.

4.2 *Airlines* dataset

A second empirical exercise is based on the *Airlines* dataset[6] [Webb et al., 2016]. This dataset is a useful test-bed for evaluating machine learning algorithms for real-world, non-stationary, streaming problems. The *Airlines* dataset consists of a large number of records, around 116 million, containing flight arrival and departure details for all commercial flights, within the U.S.A., from October 1988 to April 2008. It contains 14 variables, including the variable *Year*. In order to compute the empirical evolution of the drift over time, we split observations in respect of each year from 1988 to 2008, giving $T = 21$ temporal periods. Thus, we have a collection of datasets with the same number of $p = 13$ variables, with the variable Year removed and utilized to identify time periods. For each year $t = 1988, \ldots, 2008$, we realize graphical models with Table 8 reporting the names of the variables and the labels of the nodes. Figure 7 features graphs for selected periods suggesting that the structure of the graph changes over a period of several years. As expected from Webb et al. [2016], figure 8 depicts 3 moments of non-stationarity which correspond to the presence of drift, the first of which pertains to the period 1990 to 1993. The second drift is present in 1995 and the last in 2002. The mean of coefficient $\beta_t = -0.2$ indicates a slight drift. In this example, the dataset is more stable than that of the previous example, therefore the mean value of the intercept is $\beta_0 = 307.0$. Table 4 displays the posterior summary for the regression parameters.

4.3 *Ozone Level* dataset

The example is based on the *Ozone Level Detection* dataset [Zhang and Fan, 2008][7]. This dataset contains 2536 observations and 73 variables that include the variable *Date*, and 72 variables describing various measures of air pollutants and meteorological information for the Houston area. The period of observation covers the period from 01/1998 to 12/2004 and with this information we identify $T = 7$ periods (1998, 1999, 2000, 2001, 2002, 2003 and 2004). We group observations for each of these years and we build a panel dataset with the same variables $p = 72$ for $T = 7$. Figure 9 displays the structure of the graphs for each of the years from 1998 to 2003 showing that the behavior of the variables changes over time. Table 9 reports the names of the variables and the labels of the nodes. Figure 10 shows a constant slope of *Stability*. Despite the levels of stability being high, we note the presence of constant incremental drift. Table 5 shows the posterior summaries for the regression parameters in this example: the intercept mean $\beta_0 = 2766.30$ indicates that despite the presence of drift, the *Stability* index is high in the periods under observation, while the mean of $\beta_t = -1.4$ portrays a risk of drift over the longer term. The functionality of this example is twofold: first, it is possible to apply this approach in a somewhat high dimensional situation. Second, the approach proposed here allows for computing the drift in a dataset with variables having relationships that are easily subject to change, but without an actual significant change in the joint distribution between all variables.

[6] Source: https://moa.cms.waikato.ac.nz/datasets/
[7] Source: https://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection

4.4 *Harvard mixed gradual drift* dataset

This example is based on a dataset proposed by López Lobo [2020]. The authors have generated 20 diverse synthetic datasets (10 abrupt and 10 gradual) by using several stream generators and functions, with different numbers of features and amounts of noise. In the proposed dataset, we select that of López Lobo [2020], which belongs to the gradual dataset that comprises 41000 observations and 5 variables (X_1,X_2 and C dichotomous variable X_3 and X_4 continuous variables). We split the dataset into $T = 82$ periods, while grouping together 500 observations for each period. First, we realize the graphical model for each period t. Figure 11 shows the graphs for selected periods, specifically for $t = 1$, $t = 20$ and $t = 21$. In period $t = 22$, the minimum BIC forest has the same structure as before the occurrence of drift. Figure 12 illustrates all these considerations, in which we can observe two distinct moments of stationarity. In Table 6 are reported the posterior summaries for the regression parameters. The mean of the coefficients $\beta_t = -0.05$ notes a slight presence of drift, while the mean of the intercept $\beta_0 = 6.8$ identifies the *Stability* within the dataset. With this controlled example, we show that our model is capable of recognizing the presence of drift in a punctual manner.

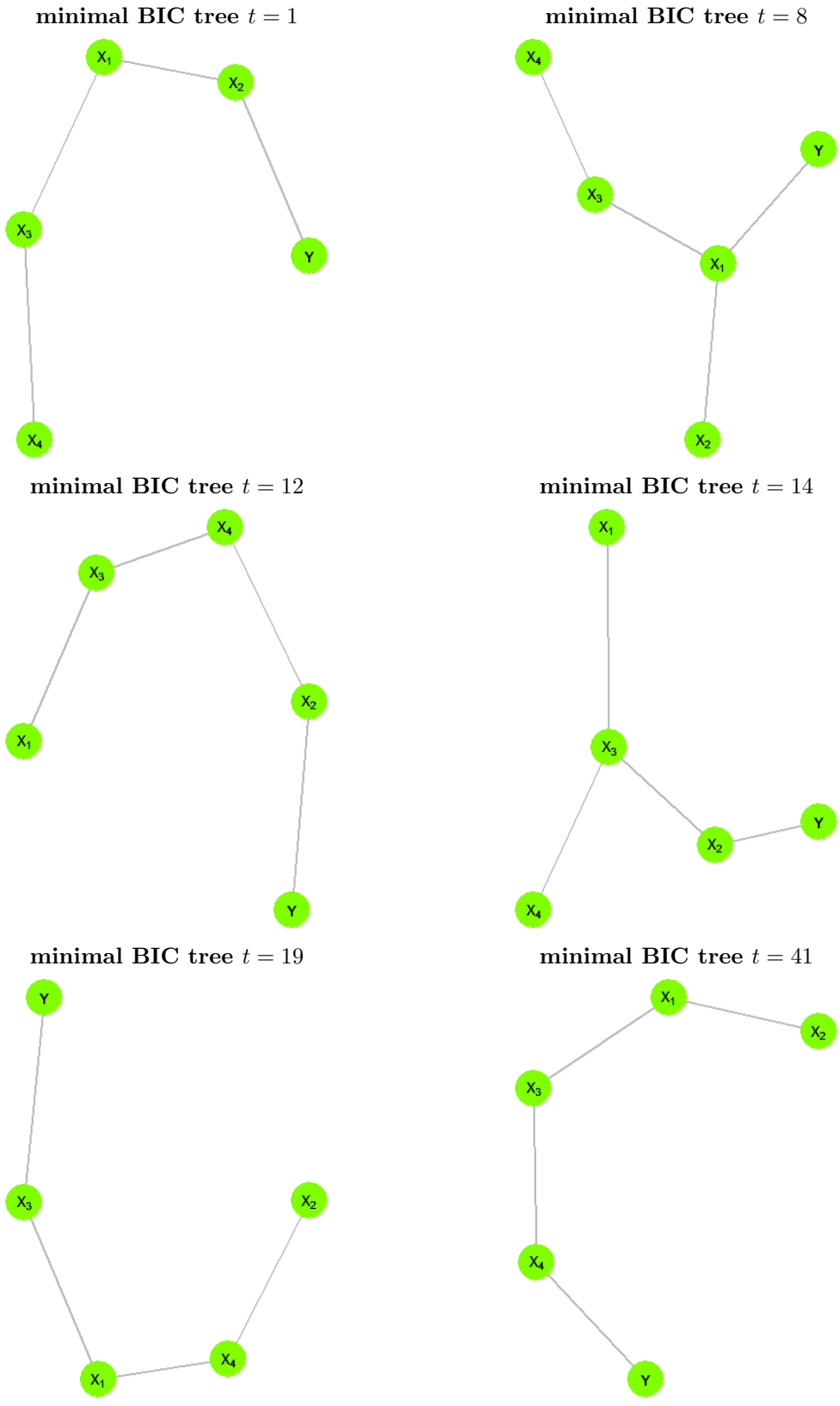

Figure 5: Evolution of the graphs, *ELEC2* dataset

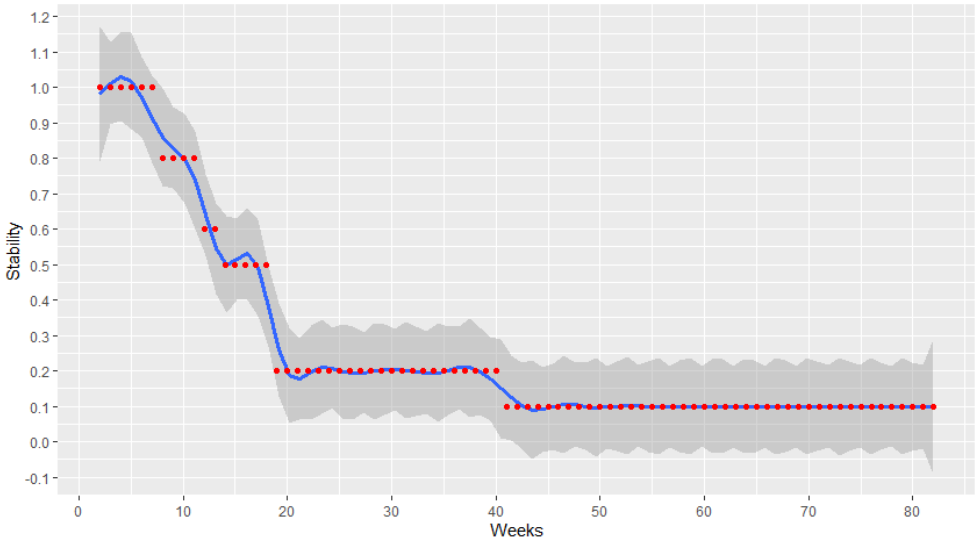

Figure 6: Evolution of Stability *ELEC2* dataset, red points represent the real values (λ_t) and the grey area indicates the credible interval

Table 3: Posterior summaries for the regression parameters, *ELEC2* dataset

Coefficients	**mean**	**s.d.**	**lower C.I.**$_{10\%}$	**upper C.I.**$_{90\%}$
β_0	7.70	0.76	6.71	8.70
β_t	-0.30	0.02	-0.32	-0.27
β_{2^T-1}	20.40	4.15	15.78	25.77

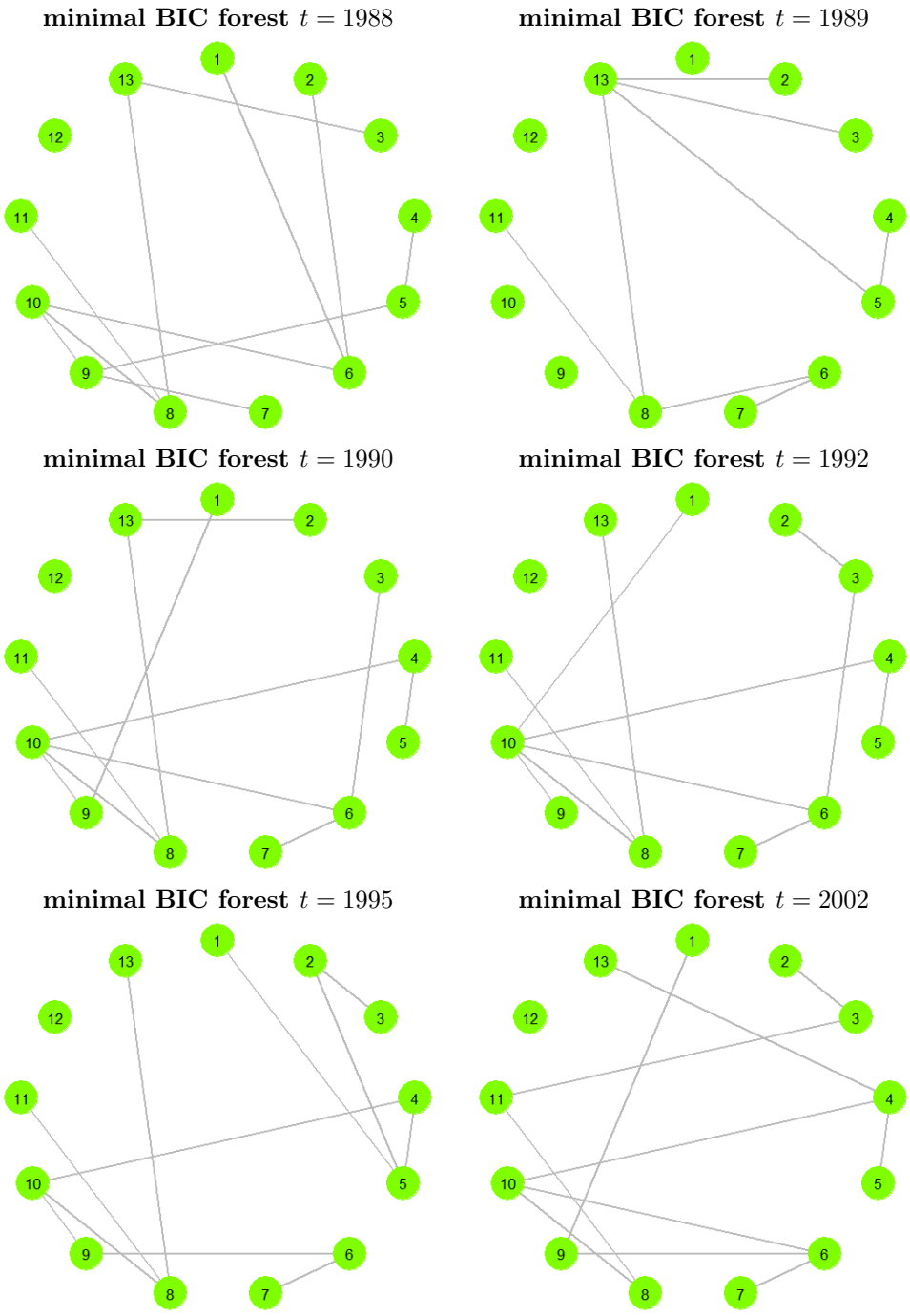

Figure 7: Evolution of the graphs, *Airline* dataset

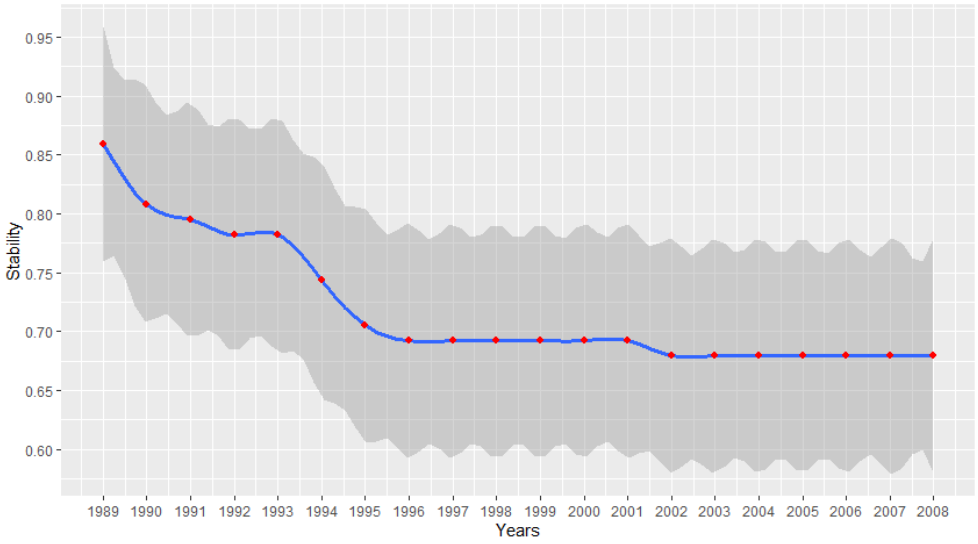

Figure 8: Evolution of Stability *Airline* dataset, red points represent the real values (λ_t) and the grey area indicates the credible interval

Table 4: Posterior summaries for the regression parameters, *Airline* dataset

Coefficients	**mean**	**s.d.**	**lower C.I.**$_{10\%}$	**upper C.I.**$_{90\%}$
β_0	307.0	73.0	214.7	400.7
β_t	-0.2	0.1	-0.2	-0.1
β_{2^T-1}	7.0	6.6	-0.5	16.3

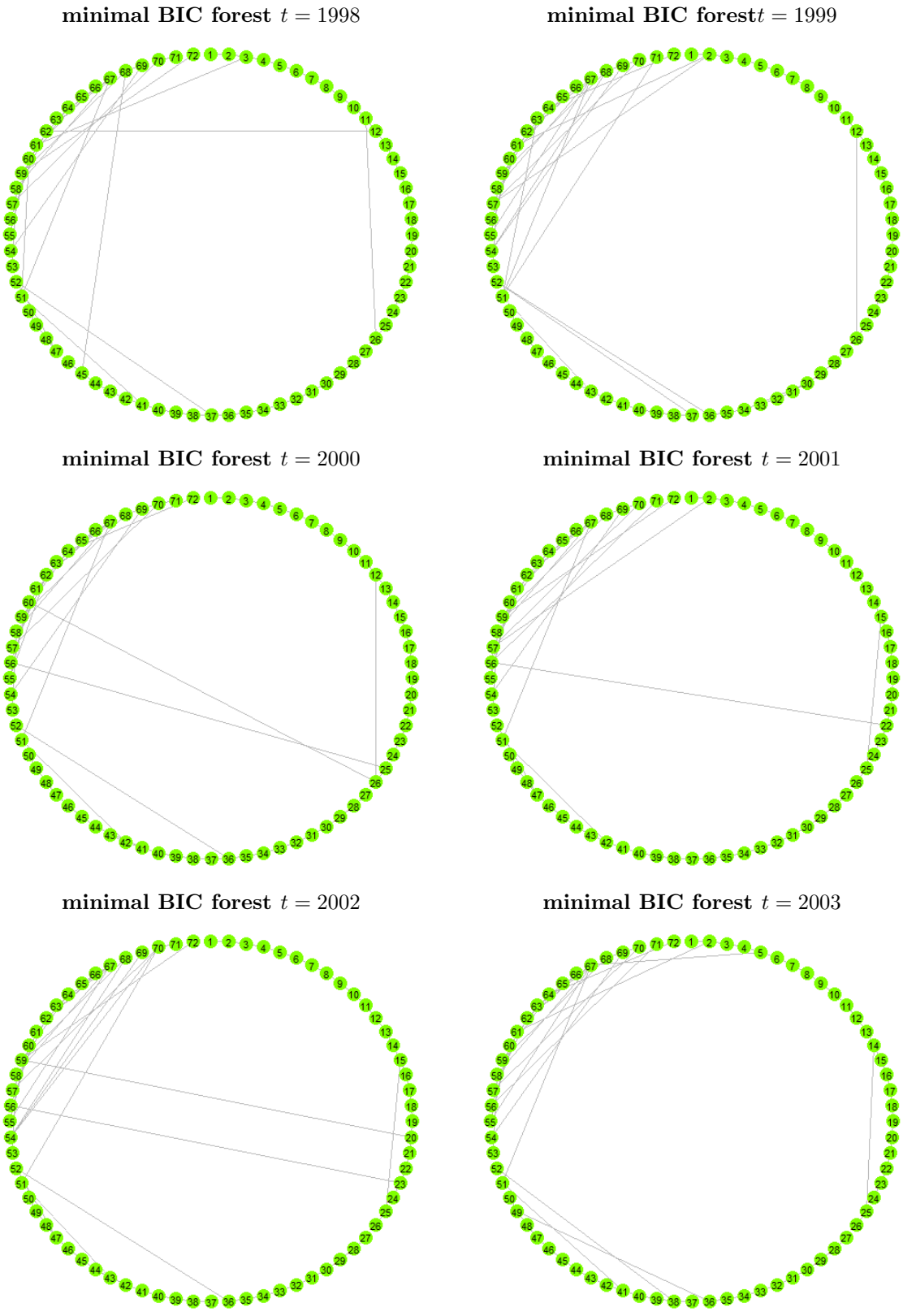

Figure 9: Evolution of the graphs, *Ozone Level* dataset

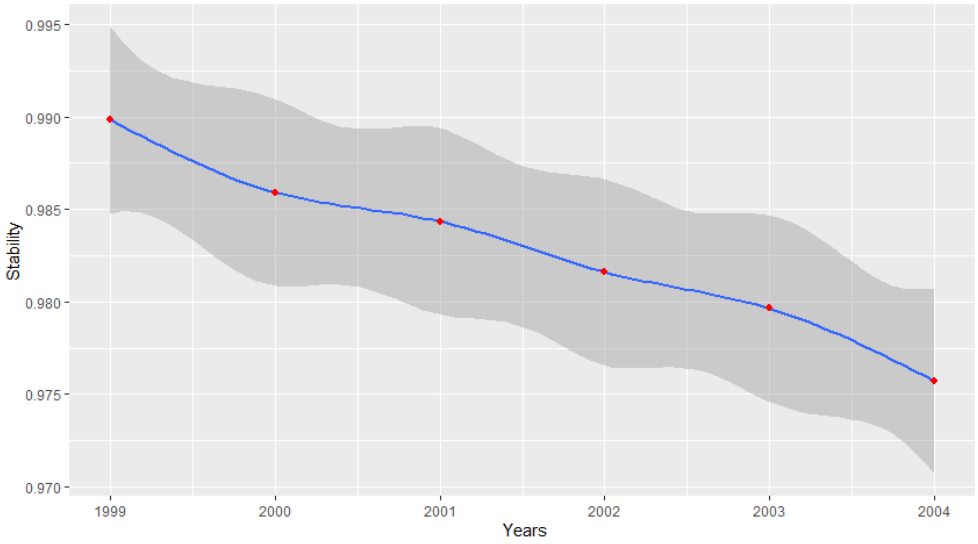

Figure 10: Evolution of Stability *Ozone Level* dataset, red points represent the real values (λ_t) and the grey area indicates the credible interval

Table 5: Posterior summaries for the regression parameters, *Ozone Level* dataset

Coefficients	**mean**	**s.d.**	**lower C.I.**$_{10\%}$	**upper C.I.**$_{90\%}$
β_0	2766.5	656.2	1960.7	2636.4
β_t	-1.4	0.3	-1.8	-1.0
β_{2^T-1}	7.0	6.5	-0.3	16.1

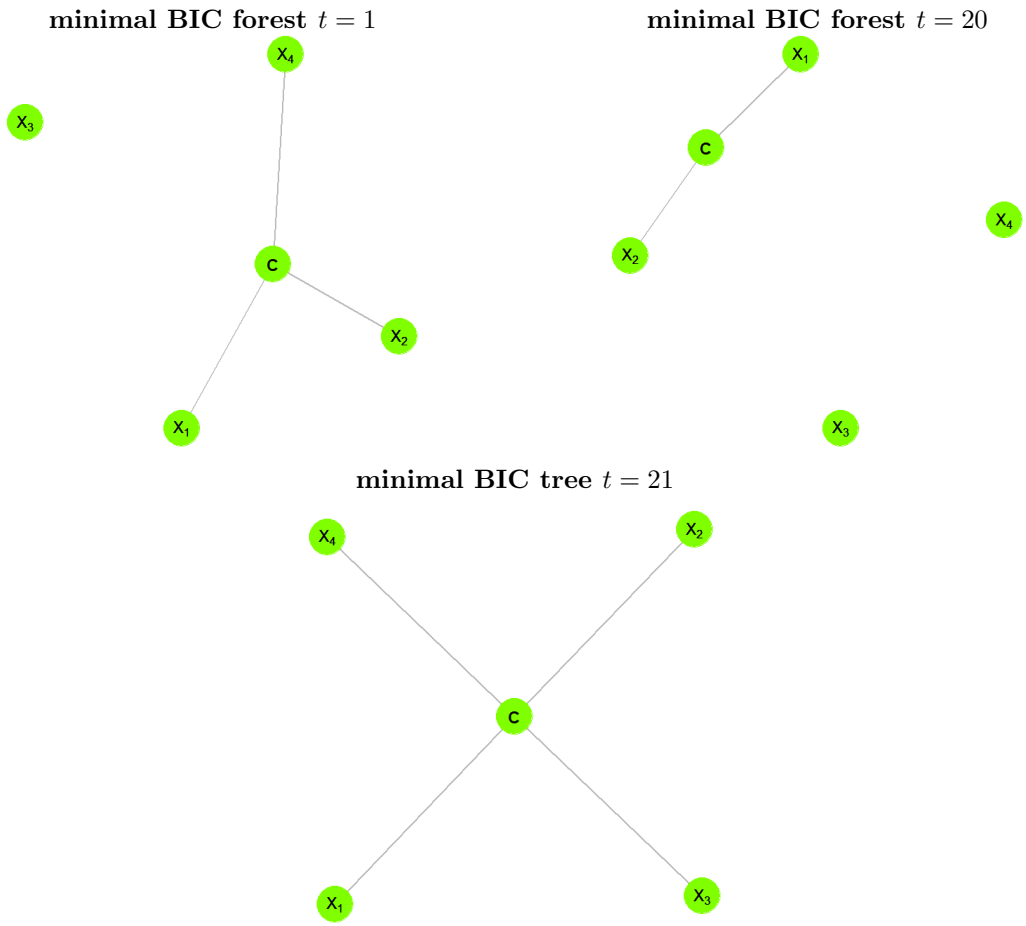

Figure 11: Evolution of the graphs, *Harvard mixed gradual drift* dataset

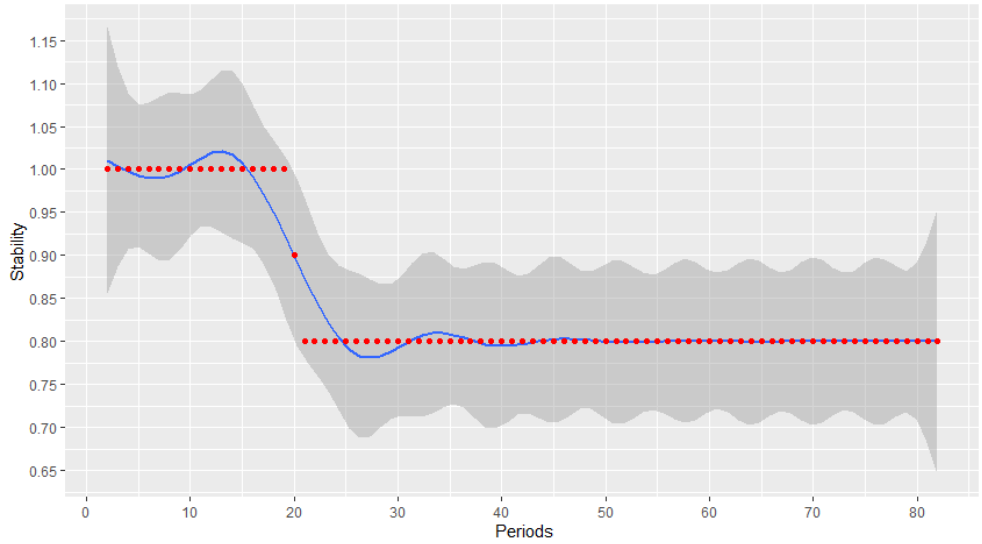

Figure 12: Evolution of Stability *Harvard mixed gradual drift*, red points represent the real values (λ_t) and the grey area indicates the credible interval

Table 6: Coefficients of logistic regression *Harvard mixed gradual drift*

Coefficients	**mean**	**s.d.**	**lower C.I.**$_{10\%}$	**upper C.I.**$_{90\%}$
β_0	6.8	1.1	5.5	8.2
β_t	-0.05	0.1	-0.1	0.1
β_{2^T-1}	7.0	6.5	-0.3	16.1

5 Conclusion

This paper has presented an algorithm for estimating the magnitude of model drift in the context of machine learning. While present solutions in relevant literature rely on the manner in which the classification errors of a specific target variable change over time, the present method describes the underlying, hidden, context with the use of graphical models, and estimates how observable context changes over time. A key step in the estimation of this underlying hidden context is the use of high dimensional graphical models as approximations of the joint distribution of the variables in the dataset. While the high dimensional graphical models make it possible to understand the overall dependency structure of discrete and/or continuous data, our intention is to investigate how these dependencies change over time. We exploit the underlying matrix representation of graphical models and we present an empirical method to measure the magnitude of the change of a graph's adjacency matrix over time. Specifically, we provide not only an assessment of the drift, which is independent from the model in use, but also an estimation of the confidence interval of this prediction. These two characteristics combined, allow us to signal when a data driven process shows an excessive risk due to drift and, therefore, needs to be retrained or re-calibrated. There are numerous possible applications such as predicting defaults, online recommendation systems, or spam filtering. More specifically, any prediction related to human behavior is prone to constant change in the data generating process, while biological and physical phenomena tend to be more stable over time. There are two limitations that arise from this application. In the first instance, although graphical models are powerful structural learning tools, we might, in the future, develop superior estimation techniques for encoding the structure of a dataset in a graph. This not withstanding, it will still be possible to apply this algorithm. The second limitation concerns the loss of information due to the multiple levels that can be assumed by the variable W. A possible extension of this work could be to conduct in-depth investigation into situations in which the links between two nodes appear and disappear. Further lines of research in this area include improvements in the estimation of different types of drift, allowing for temporary drift, and testing the index on a wider array of applications.

References

G. C. G. Abreu, R. Labouriau, and D. Edwards. High-dimensional graphical model search with the graphd r package. *Journal of Statistical Software*, 37(1):1–18, 2010. doi: 10.18637/jss.v037.i01. URL https://www.jstatsoft.org/index.php/jss/article/view/v037i01.

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

J. Albert and J. Hu. *Probability and Bayesian modeling*. CRC Press, 2019.

M. Althabiti and M. Abdullah. Classification of concept drift in evolving data stream. *Emerging Extended Reality Technologies for Industry 4.0: Early Experiences with Conception, Design, Implementation, Evaluation and Deployment*, page 189, 2020.

M. Baena-Garcıa, J. del Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavalda, and R. Morales-Bueno. Early drift detection method. In *Fourth international workshop on knowledge discovery from data streams*, volume 6, pages 77–86, 2006.

R. S. M. Barros and S. G. T. C. Santos. A large-scale comparison of concept drift detectors. *Information Sciences*, 451:348–370, 2018.

H. Borchani, A. M. Martínez, A. R. Masegosa, H. Langseth, T. D. Nielsen, A. Salmerón, A. Fernández, A. L. Madsen, and R. Sáez. Modeling concept drift: A probabilistic graphical model based approach. In *International Symposium on Intelligent Data Analysis*, pages 72–83. Springer, 2015.

R. J. C. Bose, W. M. van der Aalst, I. Žliobaitė, and M. Pechenizkiy. Handling concept drift in process mining. In *International Conference on Advanced Information Systems Engineering*, pages 391–405. Springer, 2011.

D. S. Bové and L. Held. Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410, 2011.

G. E. Box. Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, 143(4):383–404, 1980.

R. Cabañas, A. Cano, M. Gómez-Olmedo, A. R. Masegosa, and S. Moral. Virtual subconcept drift detection in discrete data using probabilistic graphical models. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 616–628. Springer, 2018.

B. P. Carlin and T. A. Louis. *Bayesian methods for data analysis*. CRC Press, 2008.

C. Carota, A. Durio, and M. Guerzoni. An application of graphical models to the innobarometer survey: A map of firms' innovative behaviour. *Department of Economics and Statistics" Cognetti de Martiis" Working Paper Series*, 2014.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

J. S. Cramer. The origins of logistic regression. *Tinbergen Institute Working Paper*, 119(4), 2002.

D. Edwards, G. C. De Abreu, and R. Labouriau. Selecting high-dimensional mixed graphical models using minimal aic or bic forests. *BMC bioinformatics*, 11(1):18, 2010.

R. Elwell and R. Polikar. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, 22(10):1517–1531, 2011.

J. Gama, R. Sebastiao, and P. P. Rodrigues. On evaluating stream learning algorithms. *Machine learning*, 90(3):317–346, 2013.

J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*, 2(4):1360–1383, 2008.

A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *technometrics*, 49(3):291–304, 2007.

T. E. Hanson, A. J. Branscum, and W. O. Johnson. Informative g-priors for logistic regression. *Bayesian Analysis*, 9(3):597–612, 2014.

M. Harries. *Splice-2 Comparative Evaluation: Electricity Pricing*. PANDORA electronic collection. University of New South Wales, School of Computer Science and Engineering, 1999. URL https://books.google.it/books?id=1Zr1vQAACAAJ.

S. S. Hussain, M. Hashmani, V. Uddin, T. Ansari, and M. Jameel. A novel approach to detect concept drift using machine learning. In *2021 International Conference on Computer & Information Sciences (ICCOINS)*, pages 136–141. IEEE, 2021.

A. Jara and T. Hanson. A class of mixtures of dependent tail-free processes. *Biometrika*, 98(3):553–566, 2011.

M. I. Jordan et al. Graphical models. *Statistical science*, 19(1):140–155, 2004.

R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American statistical Association*, 91(435):1343–1370, 1996.

R. Klinkenberg. Learning drifting concepts: Example selection vs. example weighting. *Intelligent data analysis*, 8(3):281–300, 2004.

R. Klinkenberg and T. Joachims. Detecting concept drift with support vector machines. In *ICML*, pages 487–494, 2000.

R. Klinkenberg and I. Renz. Adaptive information filtering: Learning in the presence of concept drifts. *Learning for text categorization*, pages 33–40, 1998.

L. I. Kuncheva and C. O. Plumpton. Adaptive learning rate for online linear discriminant classifiers. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 510–519. Springer, 2008.

S. Lauritzen. Graphical models, ser. *Oxford Statistical Science Series. Oxford University Press*, 1996.

E. Lesaffre and A. Albert. Partial separation in logistic discrimination. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1):109–116, 1989.

J. López Lobo. mixed˙0101˙gradual.tab. In *Synthetic datasets for concept drift detection purposes*. Harvard Dataverse, 2020. doi: 10.7910/DVN/5OWRGB/7MQTXW. URL https://doi.org/10.7910/DVN/5OWRGB/7MQTXW.

J.-M. Marin, C. P. Robert, et al. *Bayesian core: a practical approach to computational Bayesian statistics*, volume 268. Springer, 2007.

M. Nuccio and M. Guerzoni. Big data: Hell or heaven? digital platforms and market power in the data-driven economy. *Competition & Change*, 23(3):312–328, 2019.

S. M. O'brien and D. B. Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746, 2004.

S. L. Rathbun and S. Fei. A spatial zero-inflated poisson regression model for oak regeneration. *Environmental and Ecological Statistics*, 13(4):409–426, 2006.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

A. R. Syversveen. Noninformative bayesian priors. interpretation and problems with construction and applications. *Preprint statistics*, 3(3):1–11, 1998.

C. Taylor, G. Nakhaeizadeh, and C. Lanquillon. Structural change and classification. In *Workshop Notes on Dynamically Changing Domains: Theory Revision and Context Dependence Issues, 9th European Conf. on Machine Learning (ECML'97), Prague, Czech Republic*, pages 67–78. April, 1997.

G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen, and F. Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.

G. I. Webb, L. K. Lee, B. Goethals, and F. Petitjean. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, 32(5):1179–1199, 2018.

G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23(1):69–101, 1996.

K. Zhang and W. Fan. Forecasting skewed biased stochastic ozone days: analyses, solutions and beyond. *Knowledge and Information Systems*, 14(3):299–326, 2008.

P. Zhao, L.-W. Cai, and Z.-H. Zhou. Handling concept drift via model reuse. *Machine Learning*, 109(3): 533–568, 2020.

C. Zorn. A solution to separation in binary response models. *Political Analysis*, 13(2):157–170, 2005.

Appendices

A Label of nodes

Table 7: Name of the variables *ELEC2* dataset

Label Node	**Name of Variables**
X_1	*Nswprice*
X_2	*Nswdemand*
X_3	*Vicprice*
X_4	*Vicdemand*
Y	*Class*

Table 8: Name of the variables *Airline* dataset

Label Node	**Variable Name**
1	*Month*
2	*Day of Month*
3	*Day of Week*
4	*CRS Departure Time*
5	*CRS Arrival Time*
6	*Unique Carrier*
7	*Flight Number*
8	*Actual Elapsed Time*
9	*Origin*
10	*Destination*
11	*Distance*
12	*Diverted*
13	*Arrival Delay*

Table 9: Name of the variables *Ozone Level* dataset

Label Node	**Variable Name**	**Label Node**	**Variable Name**
1	*WSR0*	*37*	*T10*
2	*WSR1*	*38*	*T11*
3	*WSR2*	*39*	*T12*
4	*WSR3*	*40*	*T13*
5	*WSR4*	*41*	*T14*
6	*WSR5*	*42*	*T15*
7	*WSR6*	*43*	*T16*
8	*WSR7*	*44*	*T17*
9	*WSR8*	*45*	*T18*
10	*WSR9*	*46*	*T19*
11	*WSR10*	*47*	*T20*
12	*WSR11*	*48*	*T21*
13	*WSR12*	*49*	*T22*
14	*WSR13*	*50*	*T23*
15	*WSR14*	*51*	*T_PK*
16	*WSR15*	*52*	*T_AV*
17	*WSR16*	*53*	*T85*
18	*WSR17*	*54*	*RH85*
19	*WSR18*	*55*	*U85*
20	*WSR19*	*56*	*V85*
21	*WSR20*	*57*	*HT85*
22	*WSR21*	*58*	*T70*
23	*WSR22*	*59*	*RH70*
24	*WSR23*	*60*	*U70*
25	*WSR_PK*	*61*	*V70*
26	*WSR_AV*	*62*	*HT70*
27	*T0*	*63*	*T50*
28	*T1*	*64*	*RH50*
29	*T2*	*65*	*U50*
30	*T3*	*66*	*V50*
31	*T4*	*67*	*HT50*
32	*T5*	*68*	*KI*
33	*T6*	*69*	*TT*
34	*T7*	*70*	*SLP*
35	*T8*	*71*	*SLP_*
36	*T9*	*72*	*Precp*

B Library

The trees were build using the R library gRapHD which is available to the R community via the CRAN repository [Abreu et al., 2010].